Action Detail Matters: Refining Video Recognition with Local Action Queries

Mengmeng Wang[1] Zeyi Huang[2] Xiangjie Kong[1] Guojiang Shen[1 *]
Guang Dai[3] Jingdong Wang[4] Yong Liu[5]
[1]Zhejiang University of Technology [2]Huawei [3]SGIT AI Lab, State Grid Corporation of China
[4]Baidu [5] Zhejiang University

Abstract

Video action recognition involves interpreting both global context and specific details to accurately identify actions. While previous models are effective at capturing spatiotemporal features, they often lack a focused representation of key action details. To address this, we introduce FocusVideo, a framework designed for refining video action recognition through integrated global and local feature learning. Inspired by human visual cognition theory, our approach balances the focus on both broad contextual changes and action-specific details, minimizing the influence of irrelevant background noise. We first employ learnable action queries to selectively emphasize action-relevant regions without requiring region-specific labels. Next, these queries are learned by a local action streaming branch that enables progressive query propagation. Moreover, we introduce a parameter-free feature interaction mechanism for effective multi-scale interaction between global and local features with minimal additional overhead. Extensive experiments demonstrate that FocusVideo achieves state-of-the-art performance across multiple action recognition datasets, validating its effectiveness and robustness in handling action-relevant details.

1. Introduction

Video action recognition refers to the process of analyzing sequences of images in a video to identify and classify human actions or activities. When humans interpret actions in videos, attention is often drawn both to broad contextual changes and specific action details. This dual focus aligns with the Global vs. Local Processing theory [28] in the visual cognition field, which explains how our brains alternate between big-picture (global) and detail-oriented (local) processing, efficiently filtering out distractions to capture essential details. **Inspired by this, models should learn to capture global spatiotemporal features while focusing on action-specific details**.

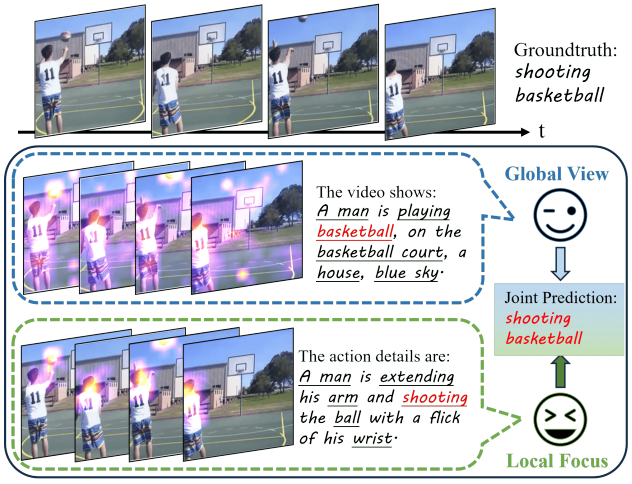

Figure 1. **High-level illustration of the proposed method.** We propose a method that combines both global view and local focuses for video action recognition. The global view captures the video's overall spatiotemporal structure, while the local focus targets action-relevant details. Integrating these two aspects allows for more accurate and refined action recognition. The highlighted areas in the visualization represent our method's attention scores, illustrating how it concentrates on key action regions.

In recent years, Transformers [6, 38] have emerged as mainstream backbone models in vision tasks, with notable adaptations in video processing [1, 2, 20, 40]. Central to Transformers is the attention mechanism, which assigns weights to sequence tokens via softmax. However, recent studies [24, 46] have shown that Transformers struggle to accurately retrieve key information due to "attention noise," which reduces the efficiency of key feature extraction and hinders performance. This phenomenon also manifests in video Transformers, where the model captures extraneous details such as background elements or unrelated objects. We visualize some examples of attention scores using a typ-

*Corresponding author: Guojiang Shen

ical video transformer [32] as shown in the global view of Fig. 1 and the first column of Fig. 5. In some cases, the model may even overlook the action's main subject altogether, resulting in misinterpretation or failure to recognize the video. While video-level global features remain essential, noise interference also highlights that **action detail matters**. Previous approaches [16, 37, 41, 42, 44, 45, 51] primarily leverage global features, leaving the representation of local, action-relevant details under-explored. Although some methods [31, 34, 48, 49] attempt to detect or represent objects, action-relevant areas are not always well-defined objects; they can be areas without clear boundaries, requiring focus on the specific area where the action actually occurs. ***Yet, how to effectively enhance action-relevant details remains challenging.***

This paper addresses the problem from two key perspectives: (i) How can we represent action-relevant areas? One possible solution could involve a separate region localization head or detection branch to identify action regions. However, this approach is resource-intensive and relies heavily on labeled region annotations which are not provided by current action recognition datasets. In fact, action details are already embedded in video features, so adaptively extracting these from global features may help bypass this issue. (ii) How can we efficiently integrate global and action-sensitive area features within a unified model? Local features need to be extracted from global features while maintaining harmony for video recognition. Achieving effective integration without excessive parameters, while balancing computational efficiency, presents a significant challenge. This requires careful design of the architecture, with a focus on controlling the number of learnable parameters to maintain model efficiency.

Keeping these two points in mind, we propose FocusVideo, a framework designed to effectively Focus on both action-relevant details and global spatiotemporal features for Video action recognition. To address the first question, we introduce learnable action queries to automatically uncover action-relevant areas within the video, inspired by recent advancements in learnable query mechanisms [3, 4, 17, 27]. Concretely, each query is initialized randomly to capture diverse aspects of actions by attending to regions associated with the action. These queries then interact with global video features, allowing the model to dynamically emphasize action-relevant areas from video representations based on the network's learning. Additionally, we use existing video features to supervise the queries, effectively bypassing the challenge of lacking region-specific labels. For the second question, we propose a local action query streaming branch and an efficient feature interaction operation, supporting the learning of action queries and enabling layerwise interactions between local queries and global video features. We utilize shared query update parameters and an autoregressive-like propagation method to iteratively refine the queries, enabling them to progressively focus on action-relevant areas. Moreover, the feature interaction can be directly and seamlessly inserted into the self-attention module of the video branch to achieve action-related feature mining and global-local integration, improving both efficiency and reducing parameter count. Extensive experiments demonstrate that FocusVideo achieves state-of-the-art (SOTA) performance on multiple action recognition datasets, validating its effectiveness and robustness.

Our main contributions can be summarized as follows:

- We propose FocusVideo, a unified framework that facilitates the integration of global and local features together, using learnable action queries to focus on action-relevant areas and suppress noise.
- We design a local action query streaming branch that allows learnable queries to progressively self-strengthen to adapt action-relevant regions, capturing omni-level action subject information and enhancing sensitivity to action details.
- We boost feature interaction efficiency with a parameter-free attention reuse strategy, integrating local query features with video features effectively.

2. Related Works

In recent years, video action recognition methods have evolved from CNN-based approaches [8, 13, 21, 39, 52] to Transformer-based approaches [1, 20, 26], adapting to changes in mainstream neural networks. Early Video Transformers focused on modifying Transformer architectures for better temporal modeling [2, 7, 10, 19], such as designing temporal input blocks or temporal Attention mechanisms. With the rise of large pre-trained models [11, 17, 23, 35, 47], methods like ActionCLIP [40] and XCLIP [29] have incorporated CLIP into video action recognition by adding temporal modules. More recent work, such as ILA [37], introduced implicit learnable alignment for better temporal modeling. Despite strong performance, these models often require costly full model fine-tuning on video data, limiting broader adoption.

To address this, Parameter-Efficient Fine-Tuning (PEFT) methods [14, 32, 33, 41, 42, 45] have emerged, aiming to adapt CLIP for video tasks by keeping most of the CLIP model frozen and only adding a few learnable parameters for efficient training. For example, EVL [22] uses a lightweight Transformer decoder on top of a fixed CLIP to achieve spatiotemporal fusion, and ST-Adapter [32], which inserts temporal adapters before attention or MLP blocks. M^2-CLIP [41] further introduces multimodal adapters and a multitask decoder for balanced supervised and generalization performance. Other methods refine image-to-video transfer from the perspective of label texts [12, 36], aiming to represent action content more clearly through language.

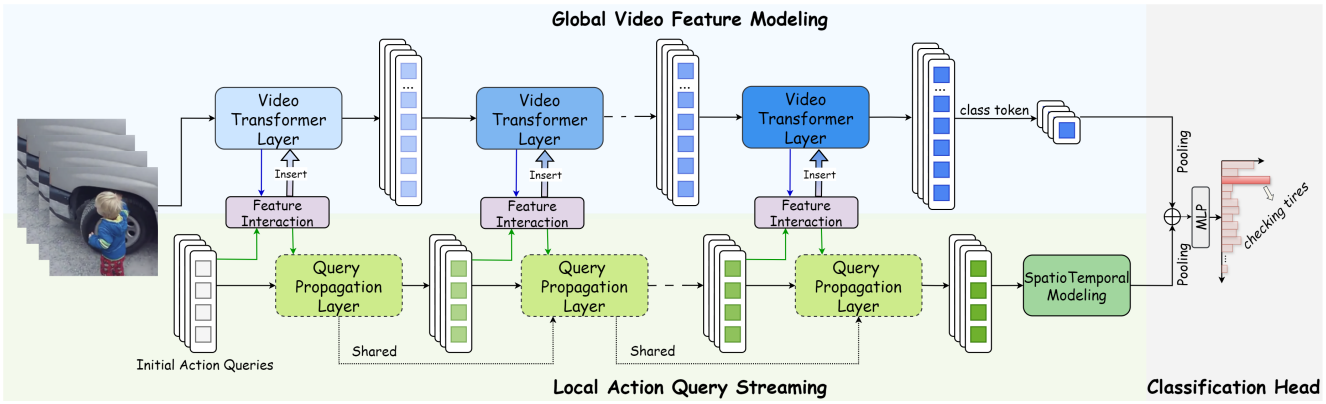

Figure 2. **Overview of FocusVideo**. The overall network architecture includes (i) a global video feature modeling branch with multiple video transformer layers to perform spatiotemporal modeling and extract video-level features, (ii) a feature interaction module for action query and video feature interaction, extracting relevant action details from the video, (iii) a local action query streaming branch with shared query propagation layers that progressively update action queries and perform spatiotemporal modeling for local action-related regions, and (iv) a classification head that integrates features from both branches (global and local) and performs the final classification.

Considering the use of auxiliary branches, we discovered that both STAN [25] and UniformerV2 [18] have incorporated these mechanisms to enhance their models. Specifically, STAN utilizes CLIP's image encoder for advanced spatiotemporal modeling, while UniformerV2 strengthens its backbone with techniques for global spatiotemporal modeling. Despite this, most of these methods focus on holistic feature learning, leaving the representation of local action-relevant regions under-explored, making them susceptible to irrelevant noise. Drawing inspiration from the Global vs. Local Processing theory [28] in visual cognition, this paper proposes an approach that enhances the model focus on critical local action regions.

3. Approach

The overall framework structure of FocusVideo is illustrated in Fig. 2. Next, we will detail the specific design of FocusVideo in this section.

3.1. Input Construction

Formally, there are two types of inputs to FocusVideo. The first is the common video frames $\mathbf{V} \in \mathbb{R}^{T\times H\times W\times 3}$, where $H \times W$ is the spatial size, and T is the number of sampled frames, which are forwarded to the global video feature modeling branch. The other is a group of learnable queries $\mathbf{A} \in \mathbb{R}^{T\times K\times C}$, representing K action-related vectors with feature channel C for T frames, allowing efficient local feature extraction without repeated image input.

3.2. Global Video Feature Modeling

FocusVideo includes a video model h_v for global spatiotemporal modeling of videos. Following the PEFT paradigm, we adopt a representative method [32] as the basis of our h_v, which effectively incorporates spatiotemporal convolution adapters inside each CLIP Transformer layer. It comprises L repeated blocks, each sequentially containing a spatiotemporal adapter (ST-AD), a multi-head self-attention (MHSA) layer, and a fully connected feed-forward network (FFN). Given the output video embedding of (l-1)-th block, $\mathbf{X}_{l-1} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\} \in \mathbb{R}^{T\times N\times C}$ of N tokens with T sampled frames, the l-th block performs the following:

$$\text{ST-AD}(\mathbf{X}_{l-1}) = \mathbf{X}_{l-1} + \text{Conv3D}(\mathbf{X}_{l-1}W_{dn})W_{up}, \quad (1)$$

$$\tilde{\mathbf{X}}_{l-1} = \text{ST-AD}(\mathbf{X}_{l-1}) + \text{MHSA}(\text{LN}(\text{ST-AD}(\mathbf{X}_{l-1}))), \quad (2)$$

$$\mathbf{X}_l = \tilde{\mathbf{X}}_{l-1} + \text{FFN}(\text{LN}(\tilde{\mathbf{X}}_{l-1})) \quad (3)$$

where LN means layer normalization and Conv3D means 3D convolution. Here, ST-AD is learnable, while the other parameters remain frozen. After L layers, the output $\mathbf{X}_L \in \mathbb{R}^{T\times N\times C}$ serves as the global video feature representation.

3.3. Seamless Feature Interaction Operation

Intuitively, to enable action queries to selectively capture local action-relevant features in a video, interaction with the video features is essential. Common interaction methods often require additional parameters, increasing the overall learning burden. To address this, we carefully devise a parameter-free feature interaction operation that can be seamlessly inserted into the video Transformer directly.

As shown in Fig. 3(a), we start by concatenating the action queries $\mathbf{A}$ with the video features $\mathbf{X}$ from each corresponding layer to form the query $\mathbf{Q}$ input of the original MHSA of the video Transformer block. Meanwhile, the key and value inputs remain as the video features. Then we

change the original self-attention mechanism in Sec. 3.2 to a cross-attention mechanism. The core attention operation can be formulated as:

$$\hat{\mathbf{X}} = \text{LN}(\text{ST-AD}(\mathbf{X})), \hat{\mathbf{A}} = \text{LN}(\mathbf{A}), \tag{4}$$

$$\mathbf{Q} = [\mathbf{Q}_X, \mathbf{Q}_A] = [\hat{\mathbf{X}}\mathbf{W}_Q, \hat{\mathbf{A}}\mathbf{W}_Q], \tag{5}$$

$$\mathbf{K} = \hat{\mathbf{X}}\mathbf{W}_K, \mathbf{V} = \hat{\mathbf{X}}\mathbf{W}_V, \tag{6}$$

$$\text{CA}([\hat{\mathbf{X}}, \hat{\mathbf{A}}], \hat{\mathbf{X}}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{7}$$

where $\mathbf{W}_Q$,$\mathbf{W}_K$ and $\mathbf{W}_V$ are the query, key, and value projection matrices, respectively. Softmax means softmax activation and $\sqrt{d_k}$ is a scaling factor. CA denotes cross-attention and $[\cdot]$ means concatenation along the token number dimension.

This operation can directly reuse the original video modeling parameters in Eq. (1-3) without introducing any additional parameters. It inherently includes both (i) the self-attention enhancement for the video features and (ii) a cross-attention interaction between the action queries and video features, which aims to extract action-sensitive local features from the video features. We achieve this by splitting Eq. (7):

$$\text{CA}([\hat{\mathbf{X}}, \hat{\mathbf{A}}], \hat{\mathbf{X}}) = \text{Softmax}(\frac{[\mathbf{Q}_X, \mathbf{Q}_A]\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \tag{8}$$

$$= [\text{Softmax}(\frac{\mathbf{Q}_X\mathbf{K}^T}{\sqrt{d_k}}), \text{Softmax}(\frac{\mathbf{Q}_A\mathbf{K}^T}{\sqrt{d_k}})]\mathbf{V} \tag{9}$$

$$= [\text{SA}(\hat{\mathbf{X}}), \text{CA}(\hat{\mathbf{A}}, \hat{\mathbf{X}})], \tag{10}$$

In Eq. (10), the first term corresponds to (i), while the second term corresponds to (ii). Note that, for simplicity, we describe the core single-head cross-attention computation in Eq. (4-10) and omit the description of the output linear layer and the layer subscript. The multi-head mechanism can be easily extended. Then, the output of video features is the same to Eq. (2-3) while the output of the action queries can be formulated as:

$$\mathbf{A}^{FI} = \mathbf{A} + \text{MHCA}(\hat{\mathbf{A}}, \hat{\mathbf{X}}) \tag{11}$$

where MHCA means multi-head cross-attention.

We seamlessly integrate the parameter-free feature interaction operation across all L video transformer layers, enabling continuous interaction between the learned queries and video features layer by layer. This approach fully leverages omni-scale video features in all layers rather than solely relying on high-level representations from the final layer, enhancing the model's capacity to capture both low- and high-level temporal details throughout the video sequence.

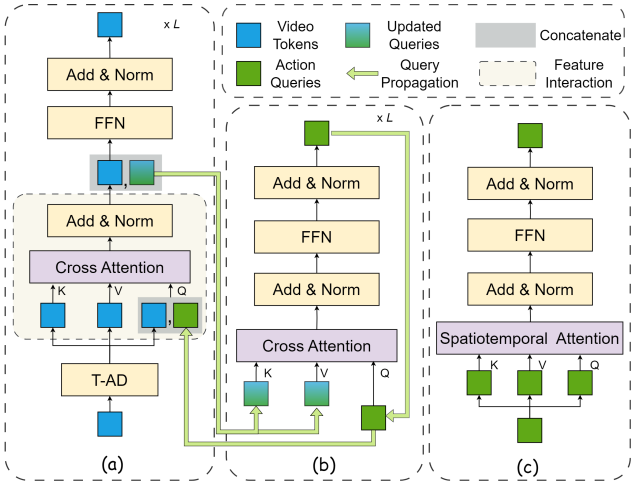

Figure 3. **Detail Structures**. (a) Global video Transformer block integrated with feature interaction module: Performs video spatiotemporal modeling and feature interaction together. "T-AD" means temporal adapter in the global branch. (b) Query propagation layer: Progressively updates action queries across layers to refine action representations. The iterative process occur iteratively across all L video transformer layers, allowing for comprehensive interaction between the learned queries and the video features at each layer. (c) Local spatiotemporal modeling: Performs spatiotemporal modeling on local action queries.

3.4. Local Action Query Streaming

Relying solely on global video representations [1, 40, 42, 51] can dilute critical spatiotemporal details associated with specific actions. Queries are designed to extract and compress core information from the input data, serving the final task [3, 17, 27]. The core idea of "query" is based on the attention mechanism, which helps the models gradually focus on specific parts of the input data. This flexible and adaptive nature of queries allows the query-based models to self-optimize their focus and effectively represent the underlying structures within the data.

Inspired by this, our method incorporates a local action query streaming branch that dynamically propagates action queries for self-enhancement across multiple feature scales within the video representation, enabling FocusVideo to accurately identify and focus on regions that are strongly correlated with the target actions, enhancing the model's attention to action-relevant areas. This branch consists of two main components: the Query Propagation (QP) layers and a local spatiotemporal modeling module, as shown in Fig. 3(b,c). The former propagates $\mathbf{A}$ by integrating the interacted information from $\mathbf{A}^{FI}$ sequentially across layers. The latter then performs spatiotemporal enhancement on all action-related queries across frames after the last QP layer.

Query Propagation Layer. When given the action queries $\mathbf{A}_j$ and the interacted queries $\mathbf{A}_j^{FI}$ at the j-th layer, the

QP layer first performs cross attention between these inputs. Here, $\mathbf{A}_j$ serves as the attention query while $\mathbf{A}_j^{FI}$ provides the key-value pairs. Notably, this operation is done frame-by-frame without inter-frame interaction because $\mathbf{A}_j^{FI}$ has already interacted with the video features containing spatiotemporal information, selecting the action-relevant spatiotemporal features from the global video branch. Then, an FFN is attached to further adjust the features. The steps can be written as:

$$\tilde{\mathbf{A}}_j = \mathbf{A}_j + \text{MHCA}(\text{LN}(\mathbf{A}_j), \text{LN}(\mathbf{A}_j^{FI})), \tag{12}$$

$$\mathbf{A}_j^{QP} = \tilde{\mathbf{A}}_j + \text{FFN}(\text{LN}(\tilde{\mathbf{A}}_j)) \tag{13}$$

To further reduce parameter counts and simplify training, we share the QP Layer parameters across all L layers, treating it as a single QP Layer applied iteratively in an autoregressive manner. Each iteration refines the representation, with the output from the previous pass serving as the input for the next, ensuring that only one set of parameters is needed throughout the entire process, thereby minimizing the number of learnable parameters and enhancing training efficiency.

Local Spatiotemporal Modeling. After obtaining the fully propagated action-related queries $\mathbf{A}_L^{QP}$, we apply a spatiotemporal self-attention mechanism to strengthen temporal and spatial dependencies between all the action queries across all frames. The features are first reshaped from $\mathbf{A}_L^{QP} \in \mathbb{R}^{T\times K\times C}$ to $\bar{\mathbf{A}} \in \mathbb{R}^{1\times TK\times C}$. We perform self-attention along the combined TK dimension as follows:

$$\tilde{\mathbf{A}} = \bar{\mathbf{A}} + \text{MHSA}(\text{LN}(\bar{\mathbf{A}})), \tag{14}$$

$$\mathbf{A}^{ST} = \tilde{\mathbf{A}} + \text{FFN}(\text{LN}(\tilde{\mathbf{A}})) \tag{15}$$

This block allows local action-related queries to capture spatiotemporal details across all frames effectively.

3.5. Classification Head

Once obtained the video feature representation $\mathbf{X}_L$ and the local action-related queries $\mathbf{A}^{ST}$, we take the class token of $\mathbf{X}_L$ and perform mean pooling along the temporal dimension to obtain the final global video features. Similarly, for $\mathbf{A}^{ST}$, we perform mean pooling over both the query dimension and temporal dimension to obtain the final local action-related features. Subsequently, we add these two features together as $\mathbf{F}$ and pass it through a Linear layer for classification.

Training Objectives. To supervise the learning of FocusVideo, we utilize two loss functions. The first is the standard classification cross-entropy loss $\mathcal{L}_{cls}$ applied to the joint feature $\mathbf{F}$ to guide the network's overall learning. In addition, we use a frame-level video feature reconstruction contrastive loss $\mathcal{L}_{\text{recon}}$ inspired by the distill loss [30, 34] to specifically guide the learning of action queries. This additional loss focuses on maximizing the alignment of action query features with the reconstructed video features. By doing so, it ensures that the query features are more consistent with the global video representation, while also guiding the queries to concentrate on areas relevant to the actions.

$$c(\mathbf{A}^R, \mathbf{X}^R) = \sum_{k=1}^{K} \frac{e^{\langle \mathbf{A}_k^R, \mathbf{X}^R\rangle/\tau}}{\sum_{l=1}^{K} e^{\langle \mathbf{A}_l^R, \mathbf{X}^R\rangle/\tau}} \langle \mathbf{A}_k^R, \mathbf{X}^R\rangle, \tag{16}$$

$$\langle \mathbf{A}_k^R, \mathbf{X}^R\rangle = \frac{\mathbf{A}_k^R \cdot \mathbf{X}^R}{\|\mathbf{A}_k^R\|\|\mathbf{X}^R\|}, \tag{17}$$

$$\mathcal{L}_{\text{recon}} = -\sum_{t=1}^{T} \log \frac{e^{c(\mathbf{A}^R, \mathbf{X}^R)/\tau}}{e^{c(\mathbf{A}^R, \mathbf{X}^R)/\tau} + \sum_{\mathbf{X}'\sim\mathcal{N}} e^{c(\mathbf{A}^R, \mathbf{X}')/\tau}} \tag{18}$$

where τ as the temperature parameter. Here, $\mathbf{A}^R$ is obtained by adding a linear layer to the spatiotemporal features $\mathbf{A}^{ST}$ to adjust region-specific features, while $\mathbf{X}^R$ represents the existing target video features for reconstruction. $\mathcal{N}$ refers to a negative sample pool consisting of other videos from the same batch that belong to different classes. $\mathcal{L}_{\text{recon}}$ maximizes similarity between matching features of $\mathbf{A}^R$ and $\mathbf{X}^R$ while minimizing the similarity to unrelated ones. The final training objective is defined as $\mathcal{L} = \lambda_1\mathcal{L}_{cls} + \lambda_2\mathcal{L}_{\text{recon}}$, optimizing both classification accuracy and action query representation quality, where λ_1 and λ_2 are the weighting factors of the two losses, respectively.

4. Experiments

4.1. Experimental Setup

We evaluate the performance of the proposed FocusVideo on two widely-used datasets: Kinetics-400 (K400) [15] and Something-Something-V2 (SSv2) [9]. We employ CLIP [35] with ViT-B/16 and ViT-L/14 as our backbones. Note that the backbones are frozen during the training process. Only spatiotemporal adapters [32] in the global branch and the whole local branch are leanable. The sparse frame sampling strategy is used with 8, 16 or 32 frames during both training and inference. The Transformer blocks of QP layer and local spatiotemporal modeling are equipped with 8 attention heads and FFNs where the hidden dimension equals the input feature size. Unless otherwise specified, our model operates with 8 action queries for every input frame. Additionally, both λ_1 and λ_2 are set to 1. The experiments are performed on four A100 for ViT-B models and eight A100 for ViT-L models.

4.2. Main Results

Results on Kinetics-400. Table 1 presents the comparisons with SOTA video models on K400 dataset. Our first observation is that the FocusVideo framework provides substantial enhancements over the baseline global video rep-

Table 1. Performance comparison on K400. The per-view GFLOPs is reported. Views mean crops×clips.

Method	TP (M)	Frames	Views	Top-1(%)	Top-5(%)	GFLOPs
MViTv2-B [20]	52	32	5 × 1	82.9	95.7	225
EVL-B/16 [22]	86	8	1 × 3	82.9	-	444
ST-Adapter-B/16 [32]	7	8	1 × 3	82.0	95.7	148
ST-Adapter-B/16 [32]	7	32	1 × 3	82.7	96.2	607
AIM-B/16 [45]	11	8	1 × 3	83.9	96.3	202
ActionCLIP-B/16 [40]	142	32	10 × 3	83.8	96.2	563
X-CLIP-B/16 [29]	132	8	4 × 3	83.8	96.7	145
Vita-CLIP B/16 [42]	39	16	4 × 3	82.9	96.3	190
STAN-conv-B/16 [25]	-	8	1 × 3	83.1	96.0	238
M^2-CLIP-B/16 [41]	16	8	4 × 3	83.4	96.3	214
M^2-CLIP-B/16 [41]	16	32	4 × 3	84.1	**96.8**	842
MoTE-B/16 [51]	-	8	4 × 3	83.0	96.3	141
OST-B/16 [5]	-	16	1 × 1	83.2	-	-
FocusVideo-B/16	15	8	4 × 3	84.1	96.5	204
FocusVideo-B/16	15	32	4 × 3	**84.7**	**96.8**	816
ST-Adapter-L/14 [32]	-	8	1 × 3	86.7	97.5	687
ST-Adapter-L/14 [32]	-	32	1 × 3	87.2	97.6	2749
AIM-L/14 [45]	38	8	1 × 3	86.8	97.2	934
DUALPATH-L/14 [33]	27	32	1 × 3	87.7	97.8	-
Text4Vis-L/14 [43]	231	32	4 × 3	87.1	97.4	1662
M^2-CLIP-L/14 [41]	54	32	4 × 3	87.0	97.6	-
MoTE-L/14 [51]	-	8	4 × 3	86.8	97.5	649
MoTE-L/14 [51]	-	16	4 × 3	87.2	97.7	1299
FocusVideo-L/14	33	8	4 × 3	87.2	97.7	914
FocusVideo-L/14	33	32	4 × 3	**88.0**	**97.9**	3656

resentation model, ST-Adapter [32]. In particular, FocusVideo achieves performance gains of 2.1% on the 8-frame setting and 2.0% on the 32-frame setting when ViT-B/16 serves as the backbone. These results demonstrate that our local action-focused branch significantly strengthens global spatiotemporal representation, directly validating the effectiveness of our approach in highlighting action-specific areas. Second, our FocusVideo with a ViT-B/16 backbone achieves 84.1% accuracy with 8-frame input, outperforming other methods using the same backbone, including fully fine-tuned models like ActionCLIP [40] and XCLIP [29], while requiring fewer learnable parameters and frames. With a 32-frame input, our method further achieves top performance, surpassing recent methods like OST [5], MoTE [51] and M2-CLIP [41]. When using the larger ViT-L/14 backbone, our FocusVideo achieves further improvements, demonstrating its scalability and effectiveness with larger architectures. These results demonstrate FocusVideo's continued efficiency and strong performance on video tasks.

Results on Something-something-v2. In Table 2, we present the performance comparisons on SSv2. Compared to the baseline global video representation model, ST-Adapter [32], FocusVideo still demonstrates a noticeable improvement, achieving a 1.3% and 1.0% performance boost when using the ViT-B/16 backbone. The extra learnable parameters added on top of ST-Adapter amount to only 8M, yet they significantly enhance the encoding of local action details, leading to a more refined and comprehensive video understanding representation. In addition, based on both ViT-B/16 and ViT-L/14, our method achieves competitive or superior performance compared to

Table 2. Performance comparison with the state-of-the-arts on SSv2. The per-view GFLOPs is reported. "F" means frames.

Model	F	Views	Top-1(%)	Top-5(%)	GFLOPs
S-ViT-B/16 [50]	16	2×3	69.3	92.1	340
ST-Adapter-B/16 [32]	8	1×3	67.1	91.2	163
ST-Adapter-B/16 [32]	32	1×3	69.5	92.6	-
ILA-ViT-B/16 [37]	8	4×3	65.0	89.2	214
ILA-ViT-B/16 [37]	16	4×3	66.8	90.3	438
AIM-ViT-B/16 [45]	8	1×3	66.4	90.5	208
AIM-ViT-B/16 [45]	32	1×3	69.1	92.2	832
STAN-B/16 [25]	16	1×3	69.5	92.7	459
Vita-CLIP-B/16 [42]	16	-	48.7	-	-
DUALPATH-B/16 [33]	32	1×3	70.3	**92.9**	-
M^2-CLIP-B/16 [41]	32	1×3	69.3	91.8	1010
OST-B/16 [5]	16	1 × 1	60.3	-	-
FocusVideo-B/16	8	1 × 3	68.4	91.0	227
FocusVideo-B/16	32	1 × 3	**70.5**	92.4	908
ST-Adapter-L/14 [32]	8	3×1	70.0	92.3	-
ST-Adapter-L/14 [32]	32	3×1	72.3	93.9	-
EVL-ViT-L/14 [22]	32	1×3	68.0	-	8086
DUALPATH-L/14 [33]	32	1×3	71.4	93.4	1932
AIM-ViT-L/14 [45]	8	1×3	67.6	91.6	959
AIM-ViT-L/14 [45]	32	1×3	70.6	92.7	3836
ILA-ViT-L/14 [37]	16	4×3	70.2	91.8	3723
M^2-CLIP-L/14 [41]	32	1×3	72.1	93.2	-
FocusVideo-L/14	8	1×3	70.7	92.3	985
FocusVideo-L/14	32	1×3	**72.9**	**94.0**	3840

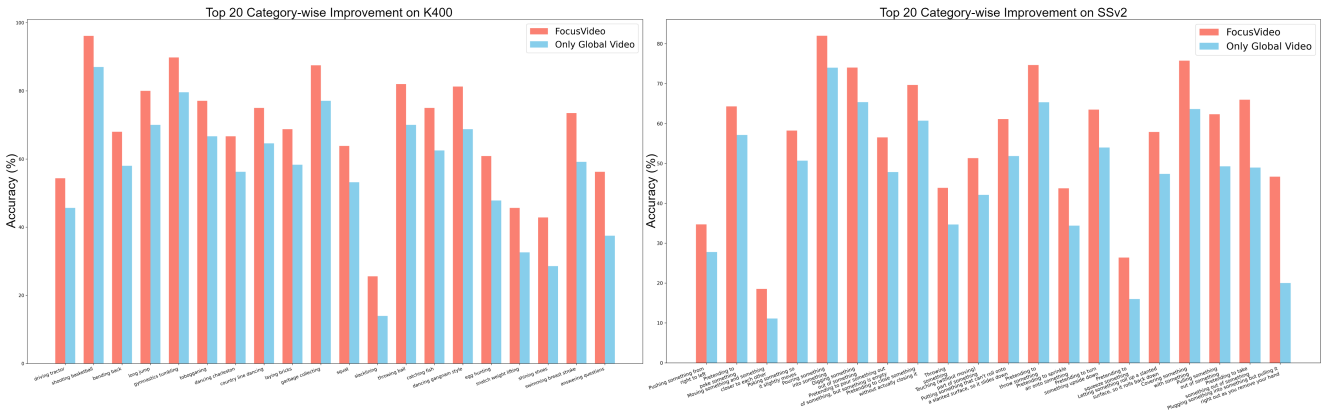

Figure 4. **Top 20 category-wise Improvements.** We visualize the top 20 categories where our method brings the most improvement compared to the pure global video modeling baseline on K400 and SSv2 datasets. The results show significant enhancements in categories that require fine-grained action details. Particularly in complex categories like "Pretending to...", highlights the importance of representing action-specific details.

most prior works. For example, compared to OST-B/16 [5], FocusVideo-B/16 outperforms it by 8.1% with 8-frame input. With 32-frame input, FocusVideo leads M2-CLIP [41] by 1.2%, while requiring fewer learnable parameters. These results demonstrate the effectiveness of our approach with the local action-focused queries, emphasizing its ability to enhance video action recognition.

4.3. Ablation and Analysis

We conduct ablation experiments on both K400 and SSv2 to validate the effectiveness of the proposed FocusVideo.

Component-wise analysis of FocusVideo. In Table 3a, we perform a detailed ablation of each proposed component in FocusVideo, gradually adding them to assess their impact. Each component is integrated with its default ablation configuration. The QP layer in M2 applies cross-attention between video features and queries from the previous layer, bypassing self-propagation. Results show that M2 notably improves performance, underscoring the value of the local branch. M3 and M4 consistently improve performance, demonstrating the effectiveness and necessity of our proposed feature interaction operation and local spatiotemporal modeling. The final M4 model achieves the highest effectiveness, solidifying it as the complete FocusVideo.

Varying Number of Action Queries. We ablate the number of action queries in Table 3b. We observe that using just 2 queries already yields good results. When increasing to 8 queries, the performance is the best, which is our final setting. However, further increasing the number of queries leads to saturation or even a slight decline in performance. We believe the reason for this is that too many queries may introduce unnecessary regions, which in turn affect the performance.

Effect of Omni-scale Propagation. In Table 3c, we ablate the omni-scale propagation of the local branch. For the usage of video features, we experiment with using only the final top layer, every other layer (half), and all layers. Additionally, the QP layer is set to either shared or non-shared configurations. Results show that using shared parameters across all layers yields the best performance with minimal parameters, demonstrating the effectiveness of multi-scale feature utilization and our autoregressive-like setup. Interestingly, when QP layer parameters are not shared, performance declines, likely due to overfitting or inconsistent feature representations across layers.

Design of Feature Interaction. We experiment with several different configurations of the feature interaction module. As shown in Table 3d, "None" indicates that this module is not used, and the video features from each layer directly replace $\mathbf{A}_j^{FI}$ in the QP layer for cross attention. The remaining three configurations represent different modules reused in the global video branch. It can be seen that reusing only the attention mechanism yields the best performance. Adding T-AD or FFN in addition to the attention mechanism imposes restrictions on local feature representation and has a negative impact on learning.

Design of Global-local Combination. We simply try three different methods to fuse the features output from the pooling of the global and local branches, as shown in Table 3e. Besides basic addition, we test concatenation followed by a linear layer to match the feature dimension, and another setting using linear projections before addition. Direct addition, which requires no extra parameters, yielded the best performance, so we selected it as the final fusion method.

Effect of Distinct Query Supervision. In Table 3f, we experiment with different supervision signals for the local branch. "None" represents using only the classification loss $\mathcal{L}_{cls}$ without additional losses. "Cls Cross entropy" adds

Table 3. Ablation studies with 8-frame FocusVideo-B/16, reporting Top-1 accuracy on K400 and SSv2. Default settings are in gray.

(a) Component-wise analysis of FocusVideo.

Models	Configuration	K400	SSv2
M1	Global Video Branch	82.4	66.8
M2	M1 + QP Layer	83.5	67.7
M3	M2 + Seamless Feature Interaction	83.9	68.0
M4	M3 + Local Spatiotemporal Modeling	**84.1**	**68.4**

(b) Varying Number of Action Queries.

Numbers	K400	SSv2
2	83.8	67.9
4	83.8	68.1
8	**84.1**	**68.4**
16	84.0	68.2

(c) Effect of Omni-scale Propagation.

Type	Local Params	K400	SSv2
Only top layer	7.9M	83.7	67.7
shared half layers	7.9M	83.9	68.0
shared all layers	7.9M	**84.1**	**68.4**
Non-shared all layers	46.9M	83.4	66.5

(d) Design of Feature Interaction.

Type	K400	SSv2
None	83.8	68.0
Attention	**84.1**	**68.4**
Attention + T-AD	82.7	67.2
Attention + FFN	83.2	67.7

(e) Design of Global-local Combination.

Type	K400	SSv2
Concatenation + Linear	83.6	67.9
Add	**84.1**	**68.4**
Linear + Add	83.9	68.3

(f) Effect of Distinct Query Supervision.

Type	K400	SSv2
None	82.7	67.0
Cls Cross entropy	83.5	67.4
Reconstruction (online)	83.4	67.2
Reconstruction (offline)	**84.1**	**68.4**

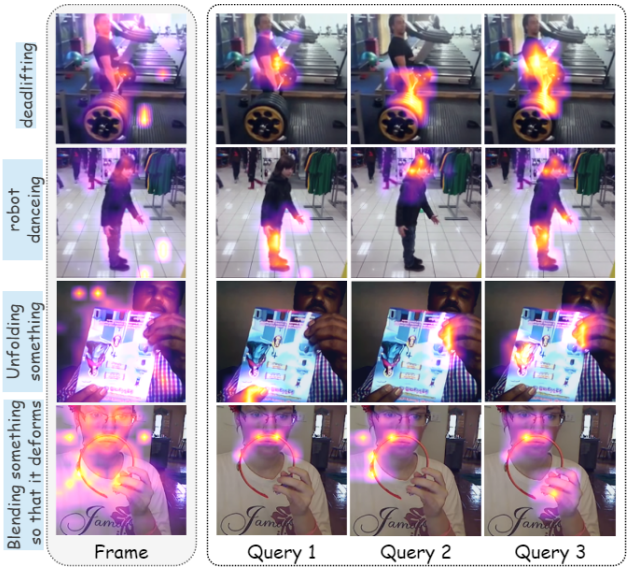

Figure 5. **Attention score visualizations.** The first column represents the attention score of the entire frame from the class token of the global video representations. The subsequent three columns show the attention score distributions of three action queries across the entire frame, indicating how each query focuses on different parts of the video content.

a linear projection to the pooled $\mathbf{A}^{ST}$, followed by a classification cross-entropy loss supervised with the video labels. "Reconstruction" uses the $\mathcal{L}_{\text{recon}}$ described in Sec. 3.5. "online" and "offline" respectively use synchronous video features and off-the-shelf trained video features for supervision. Results show that offline supervision with independently trained features performs best, helping compensate for the absence of ground truth action-region labels.

4.4. Visualization

To illustrate the effectiveness of our action queries, we visualize attention score distributions in Fig. 5. In the holistic video features shown in the first column, attention often disperses across the entire frame, capturing irrelevant areas like backgrounds or unrelated people in meanwhile. For instance, in the second row (robot dancing), attention covers the ground and bystanders rather than focusing on the action. In contrast, our action queries concentrate on crucial areas, such as the dancer's head, hands, and legs. It achieves finer-grained focus like hand-object contact in the third and fourth rows. Additionally, although we didn't explicitly enforce diversity among the queries, they effectively capture different regions, even when overlapping spatial areas. This indicates that each query can focus on distinct aspects of the action, enhancing the model's understanding of complex action patterns within overlapping yet specialized attention zones.

Moreover, to better understand the impact of adding our designed local branch, we visualize the top 20 categories where our proposed FocusVideo model shows the most improvement compared to using only global video features in Fig. 4. We can observe that our method consistently excels in classes that demand attention to fine action details across both datasets like "dancing gangnam style", "answering question" and "pretending to turn something upside down". This enhancement suggests that the local branch better captures intricate action details, contributing to higher accuracy in detail-sensitive categories.

5. Conclusion

This paper demonstrates a unified framework for integrating global context with action-focused details, yielding state-of-the-art performance in video action recognition. Our success stems from: (i) the introduction of learnable action queries, which enable our model to effectively capture critical regions in videos while filtering out irrelevant noise; (ii) the propagating local query branch, which progressively self-enhances and efficiently integrates with the global branch by sharing parameters across layers, reducing computational costs; and (iii) a parameter-free feature interaction strategy, which ensures effective interaction between global and local features across omni-scale layers without excessive computational burden.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62403429, No. 62476247, Zhejiang Provincial Natural Science Foundation of China under Grant No. LQN25F030008.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6836–6846, 2021. 1, 2, 4

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, pages 813–824, 2021. 1, 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2, 4

[4] Ting Chen and Lala Li. Fit: Far-reaching interleaved transformers. *arXiv preprint arXiv:2305.12689*, 2023. 2

[5] Tongjia Chen, Hongshan Yu, Zhengeng Yang, Zechuan Li, Wei Sun, and Chen Chen. Ost: Refining text knowledge with optimal spatio-temporal descriptor for general video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18888–18898, 2024. 6, 7

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 1

[7] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, pages 6824–6835, 2021. 2

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 2

[9] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, pages 5842–5850, 2017. 5

[10] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *CVPR*, 2022. 2

[11] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *CoRR*, abs/2102.05918, 2021. 2

[12] Chengyou Jia, Minnan Luo, Xiaojun Chang, Zhuohang Dang, Mingfei Han, Mengmeng Wang, Guang Dai, Sizhe Dang, and Jingdong Wang. Generating action-conditioned prompts for open-vocabulary video action recognition. In *ACM Multimedia 2024*, 2024. 2

[13] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. Stm: Spatiotemporal and motion encoding for action recognition. In *ICCV*, pages 2000–2009, 2019. 2

[14] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV*. Springer, 2022. 2

[15] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5

[16] Dongho Lee, Jongseo Lee, and Jinwoo Choi. Cast: Cross-attention in space and time for video action recognition. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 4

[18] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022. 3

[19] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. In *ICLR*, 2022. 2

[20] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022. 1, 2, 6

[21] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019. 2

[22] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. *arXiv preprint arXiv:2208.03550*, 2022. 2, 6

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[24] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173, 2024. 1

[25] Ruyang Liu, Jingjia Huang, Ge Li, Jiashi Feng, Xinglong Wu, and Thomas H Li. Revisiting temporal modeling for clip-based image-to-video knowledge transferring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6555–6564, 2023. 3, 6

[26] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Pro-*

ceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3202–3211, 2022. 2

[27] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020. 2, 4

[28] David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, 9(3): 353–383, 1977. 1, 3

[29] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, 2022. 2, 6

[30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5

[31] Yangjun Ou, Li Mi, and Zhenzhong Chen. Object-relation reasoning graph for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20133–20142, 2022. 2

[32] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022. 2, 3, 5, 6

[33] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Dual-path adaptation from image to video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2203–2213, 2023. 2, 6

[34] Rui Qian, Shuangrui Ding, and Dahua Lin. Rethinking image-to-video adaptation: An object-centric perspective. In *European Conference on Computer Vision*, pages 329–348. Springer, 2024. 2, 5

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2, 5

[36] Hanoona Rasheed, Muhammad Uzair khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Finetuned clip models are efficient video learners. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[37] Shuyuan Tu, Qi Dai, Zuxuan Wu, Zhi-Qi Cheng, Han Hu, and Yu-Gang Jiang. Implicit temporal modeling with learnable alignment for video recognition. *arXiv preprint arXiv:2304.10465*, 2023. 2, 6

[38] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1

[39] Mengmeng Wang, Jiazheng Xing, Jing Su, Jun Chen, and Yong Liu. Learning spatiotemporal and motion features in a unified 2d network for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3347–3362, 2022. 2

[40] Mengmeng Wang, Jiazheng Xing, Jianbiao Mei, Yong Liu, and Yunliang Jiang. Actionclip: Adapting language-image pretrained models for video action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1, 2, 4, 6

[41] Mengmeng Wang, Jiazheng Xing, Boyuan Jiang, Jun Chen, Jianbiao Mei, Xingxing Zuo, Guang Dai, Jingdong Wang, and Yong Liu. A multimodal, multi-task adapting framework for video action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5517–5525, 2024. 2, 6, 7

[42] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Vita-clip: Video and text adaptive clip via multimodal prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23034–23044, 2023. 2, 4, 6

[43] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Revisiting classifier: Transferring vision-language models for video recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2847–2855, 2023. 6

[44] Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6620–6630, 2023. 2

[45] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. *arXiv preprint arXiv:2302.03024*, 2023. 2, 6

[46] Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. *arXiv preprint arXiv:2410.05258*, 2024. 1

[47] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *CoRR*, abs/2111.11432, 2021. 2

[48] Ce Zhang, Changcheng Fu, Shijie Wang, Nakul Agarwal, Kwonjoon Lee, Chiho Choi, and Chen Sun. Object-centric video representation for long-term action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6751–6761, 2024. 2

[49] Yubo Zhang, Pavel Tokmakov, Martial Hebert, and Cordelia Schmid. A structured model for action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9975–9984, 2019. 2

[50] Yucheng Zhao, Chong Luo, Chuanxin Tang, Dongdong Chen, Noel Codella, and Zheng-Jun Zha. Streaming video model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14602–14612, 2023. 6

[51] Minghao Zhu, Zhengpu Wang, Mengxian Hu, Ronghao Dang, Xiao Lin, Xun Zhou, Chengju Liu, and Qijun Chen. Mote: Reconciling generalization with specialization for

visual-language to video knowledge transfer. *arXiv preprint arXiv:2410.10589*, 2024. 2, 4, 6

[52] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 695–712, 2018. 2